EPiC Series in Computing

Volume 105, 2025, Pages 214–220

Proceedings of EUNIS 2024 annual congress in Athens

Process-oriented approach to develop research data long-term retention service in the research institution

Minna Harjuniemi[1] and Ville Tenhunen[1]
[1] University of Helsinki, Finland
minna.harjuniemi@helsinki.fi, ville.tenhunen@helsinki.fi

Abstract

This article presents a process-oriented way to develop long-term retention of research data as a university-level service. In addition to the processes, the article examines the service management model, in which several university units participate. In addition, a straightforward technical solution is presented, which offers service development both horizontally and vertically in the future.

1 Introduction

It is quite usual that researchers ask ”where can I save this valuable data for future use or as evidence of research already completed?” Correspondingly, it is customary to consider the long-term storing of research data as part of the data life cycle and end up solving the mentioned question as an archiving question.

However, it is the case that universities and other research performing organizations have different needs to keep data for longer periods of time, not just datasets produced for articles after the research has ended.

The University of Helsinki decided to strengthen its research data storage services and data curation over the next five years. The upcoming Databank service differs from a conventional data storage service. The service developed for this is intended for inactive data that is not in active processing and is not necessarily suitable to be transferred to the service for the national long-term storage of research data Fairdata PAS (Ministry of Education and Culture, n.d.). The data do not have to be published or linked to a publication.

There are numerous use cases where data might be valuable enough to be stored. Here are some examples of the potential use of the new service:

R. Vogl, L. Desnos, J.-F. Desnos, S. Bolis, L. Merakos, G. Ferrell, E. Tsili and M. Roumeliotis (eds.), EUNIS 2024 (EPiC Series in Computing, vol. 105), pp. 214–220

- Scientific devices produce raw data which is worth storing some years before they are analyzed with new technologies or in the next research project (which will be funded later).
- A large survey which will be partially used for several projects or publications. The base data of the survey might be valuable enough to store over a long period to create time series, etc.
- Data from a research project which will end and whose output must be stored somewhere due to the funder's rules.
- Data from an external source which will be used in research projects years later.

The University of Helsinki decided to develop the Databank (University of Helsinki, 2024) service in collaboration with the owners of the main processes of research data retention and storing. Concept and main processes, service management model and technical implementation have been presented in the following chapters.

The service is still in its early stages, so systematic user feedback or their evaluations are not yet available.

2 Concept and processes

2.1 Concept design

The concept design characteristics of the Databank are

- Long-term storing service in the University of Helsinki (up to 15 years)
- Retention service and storage for non-active data
- Data which has been assessed valuable enough to store in the Databank (faculty level decisions)
- Faculties have their own quota they use when they make value assessment
- Process to select data to store in Databank i.e. check if the data applicable for the national long term preservation system (PAS), data curation services, legal assessment etc.
- Capacity for the service will be increased step by step along needs (first step ~900TB). Most of it is funded university level, but faculties can add their quota by financing components to the same system
- Implement capacity services as a service i.e. user interface remains the same while back-end capacity will be updated.
- Service management and governance processes are applications of the FitSM (ITEMO, 2024) standard principles.

According to the concept of the Databank, stored datasets possibly come from various phases of the research data management processes such as data capturing, immediate storing, data wrangling, data analysis, curation, opening, or data sharing as presented in the Figure 1. The figure shows generic research data management processes with data mapping method (Tenhunen & Wilson, 2020).

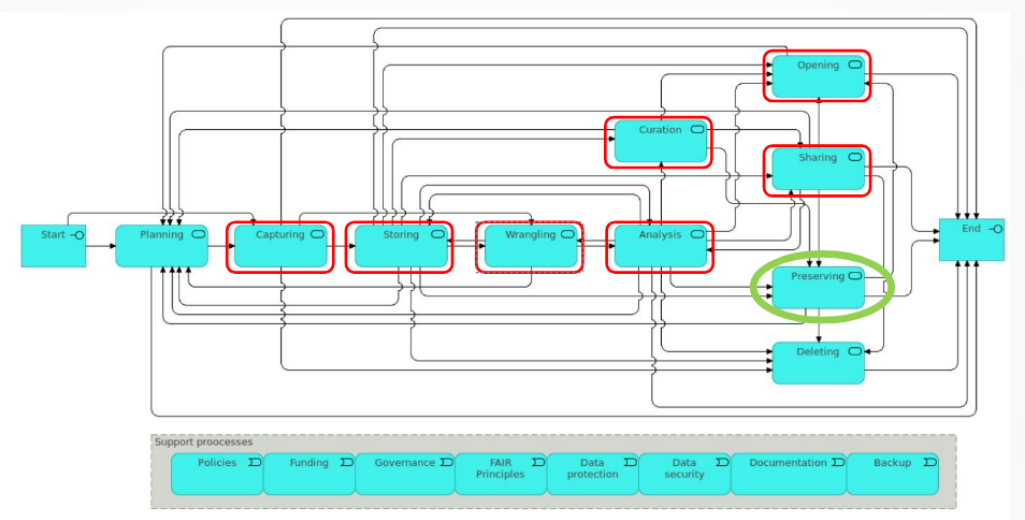

Figure 1 RDM processes and sources of datasets to the Databank

2.2 Processes of the service

The Databank service processes have several actors from several departments of the university because the whole service process requires a lot of various expertise. The process is presented in Figure 2.

- Main sub-processes are:
  - Service request and request review, operated by the data team of the University Library
  - Proposal for the research committee of the faculty
  - Assessment by the research committee of the faculty
  - Descriptive metadata management by the university library
  - Data transfer and storing by the Center for the Information Technology of the university (capacity services)
  - Data maintenance operated by the data team of the University Library
  - Annual review by the faculty research committee

Each of the faculties have a research committee, that among other things prepares guidelines and development measures concerning Faculty's research. Evaluation of data value will be a new role for these committees, and methods and criteria will develop during coming months and years.

Also, the Data Support service of the university has its role in the Databank process because it is responsible for cases which must move out from the Databank main process. The Data Support service is also usually the first contact point and promotes the service as a part of the research data service catalogue.

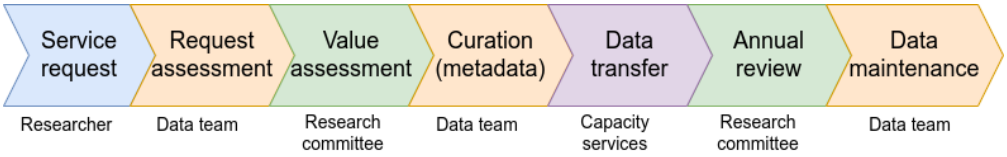

Figure 2 The Databank management process phases

3 Service management model and terms of use

3.1 Service management model

The service management system (SMS) of the Databank has some features from the FitSM standard. The service management model contains general generic SMS levels i.e. Governance level, control level and operational level.

Role model for the SMS have 4 major roles:

- Service owner; overall responsibility for a service and its definitions
- Process owners, overall accountability for processes including process goals and monitoring. Process owners also have authority to provide resources for the process
- Process managers; Responsibility for operational effectiveness and efficiency of processes
- Process staff members; Responsible for performing a specific process activity

At the University of Helsinki, the owner of the service is the university library. The process owners and managers are in the library and the IT center, where the teams also work. The university library, IT center and research administration of the university are partners in the Steering Committee of the service.

Main documents of the SMS in this context are Service Level Agreement which describe the service for users and Operational Level Agreement where process owners and has described resources and service levels to the service owner.

Basic idea is not to implement the whole FitSM standard as such but select some main features to ensure systematic service management and its continuous development in the multi-organizational environment where traditions and working cultures are different.

3.2 Terms of use

Research data is stored in the service longer than many individual research projects last. In addition, the university also has an interest in keeping materials it has determined to be valuable for a long time, for example after the researcher who created the data has retired or otherwise left the university.

That is why the university must have the right to edit and process data in the Databank. Correspondingly, the university's researchers and organizations have obligations that must be fulfilled for the data to be included in the Databank.

The units i.e. faculties are offered a certain quota of storage capacity from the service in proportion to their total budget. The unit decides whether to take the storage capacity into use and at the same time agrees to the terms of use of the service.

Users of the service, both units and researchers, commit to a set of terms of use, which ensures appropriate use of the storage capacity and the necessary documentation and metadata for the data sets

As part of the service provision process, the research committees of the units decide which data can be stored in the unit's quota. If needed, they can suggest to faculty to increase storage capacity with faculty funding.

The selection phase involves reviewing the data and deciding whether there are grounds for storing it. Following questions are examples of topics in the review:

- Is the data unique, can the same data be collected again?
- Is it part of a time series?
- Are there legal grounds?
- Does the funder or publisher require storage/preservation?
- Is there a known future project or project idea where the data could be used?
- Is there any other reason?

All users must accept the terms of use, even though employees at the university already have different obligations in their employment contracts. However, they contain various terms because different agreements have been made in different decades. Key issues of the terms of use are, for example:

- co-rights are granted to the university i.e. university needs co-rights to manage data sets also in case where researcher is not member of the university community
- documentation related to the data is attached to the service request.
- the contact person information for the data and an alternate
- the researcher undertakes to provide the minimum metadata required for the data

Without the minimum metadata, the data cannot be deposited, the units cannot take different decisions on this for their own shares of the capacity. The metadata collected in the application will be stored into the upcoming metadata catalogue which makes also data publishing service possible.

4 Technical implementation

Figure 3 presents an implementation of the Databank as level 2 (Containers) diagram of the C4 model (Brown, 2024). The figure express users of the system (dark blue), internal components of the systems (blue), external systems used for the service (grey) and their relations (arrows).

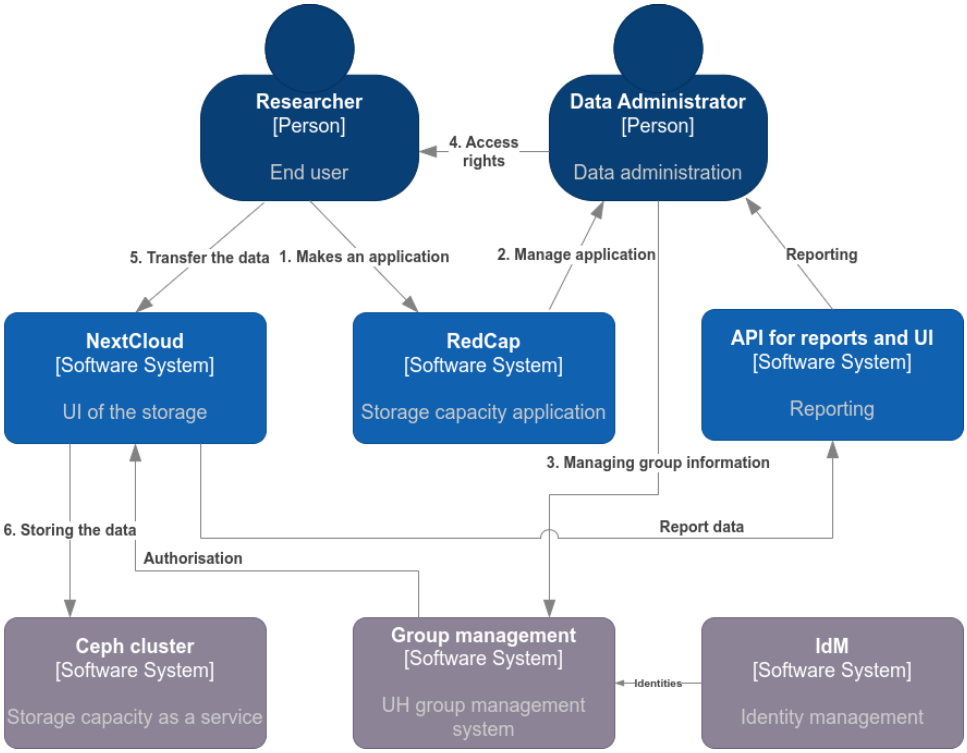

Figure 3 An implementation of the Databank as level 2 (Containers) diagram of the C4 model

Main users are researchers who want to store the data to the Databank and group of data administrators who curate, administrate and maintain the data in the Databank.

The workflow contains following main parts;

- Researcher makes service request in the RedCap form. RedCap is the web application to managing online surveys
- Data administrators (i.e. the data teams) manage the request and all processes of decision making etc.
- If requested quota has granted, data admins add researcher(s) to the specified group in the university's group management system needed for authorization of the system.
- Data administrators give also access rights to the system and inform the researcher about method to transfer data to the Databank.
- When grated, the researcher transfers the data to the Databank by using NextCloud application which stores it to the University's Ceph based object storage environment.
- For units of the university data administrators create also reports that units have correct information about their quota usage.

The technical implementation is made of separate components so that in the future it will be possible to replace individual components without having to replace the entire system at once. It is also needed that the capacity of the system can be increased by adding components as needed.

5 Conclusions

The concept of the Databank is valuable for researchers but also for university. Researchers get a chance to store systematically the research data from several stages of the data lifecycle and processes. The possible use cases of raw data may not be known when a research project is finished, and a service like Databank still enables its innovative use for a longer time. University gets more information about the data it manages and holds. At the end of the day, the data is one of the most important research outcomes of the university.

The Databank modular solution gives possibilities to improve the implementation of the FAIR principles (Wilkinson, 2016) in the university. Implementation of FAIR principles is an essential part of modern research management.

The process approach in service management gives possibilities to improve methodology and makes multi-organizational service easier. When focus is in workflows and procsess, not in separate applications or infrastructure components, overall integrations are possible to achieve in sustanable way.

In the first version of the Databank metadata management is based on manual operations. Currently documentation is on the RedCap but will be moved to the metadata catalogue in the future. Also monitoring and reporting relies on manual data collection. This should be automated in the future.

Even if the basic technical implementation is not suitable for highly sensitive research data, the process to manage data in the Databank fulfills requirements for also sensitive data. Therefore, it is reasonable to consider adding also new types of research data to Databank process.

Clarifying processes, agreements and university regulation regarding research data helps each participant understand their responsibilities in data lifecycle management.

References

Brown, S. (2024, February 7). *Container diagram*. Retrieved from The C4 model for visualising software architecture: https://c4model.com/

ITEMO. (2024, February 7). *Standards for lightweight IT service management*. Retrieved from Standards for lightweight IT service management: https://www.fitsm.eu/

Ministry of Education and Culture. (n.d.). *Digital Preservation Service for Research Data*. Retrieved from Fairdata.fi: https://www.fairdata.fi/en/dps-for-research-data/

Tenhunen, V., & Wilson, J. A. (2020, May 20). *The Processes Behind Research Data Management.* Retrieved from EUNIS; Best Paper Award: https://www.eunis.org/wp-content/uploads/2020/06/80-Tenhunen_Wilson_eunis2020_full_paper_final_2020-05-15.pdf

University of Helsinki. (2024, February 7). *University of Helsinki Databank*. Retrieved from The Data Support at the University of Helsinki: https://www.helsinki.fi/en/research/services-researchers/data-support/preservation-research-data/databank

Wilkinson, M. D. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data 3:160018 doi: 10.1038/sdata.2016.18* .